UC-6 Covid-19 Data Analysis Using Linear Regression

Noah Druss

Abstract

Covid-19 is the most impactful event in the past century. This study focuses on using different types of linear regression to predict and analyze Covid-19 infection and death data based on different features. First, simple and polynomial regression were used to predict deaths based on infections. It was found that polynomial regression at a higher degree was superior for each country with R^2 scores as high as .9998. Next linear regression coefficients were used to analyze 18 countries based on four features. When these countries were organized by wealth of the continent, death rate, and infection rate, it was found that GDP and Stringency Index were negatively correlated with infection rate.

Introduction

The key point and purpose to this study is to effectively analyze Covid-19 data by answering the research questions below. Linear regression models combined with data preprocessing was used to perform this analysis on the given dataset of over 60,000+ lines. The models used in this study are simple linear regression, multiple linear regression, and polynomial regression.

Research Question(s)

1. Can multiple linear regression find coefficients for each feature provided?
2. How accurately can Covid-19 deaths be predicted given infection data with simple linear regression and polynomial regression?
3. How significantly will these models' performance change at a country level vs global level?
4. Can linear regression be used with specific countries to elicit patterns?

Materials and Methods

The materials used in this study included the Our World in Data Covid-19 dataset and the Python libraries listed in the Tools section. The OWID dataset is a 60,000+ line CSV file with over 50 features that pulls information from the Johns Hopkins University's Covid-19 data repository.

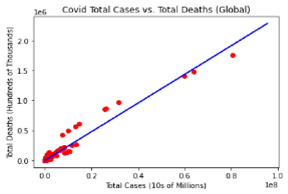
The methods used in this study included simple linear regression, multiple linear regression, and polynomial regression. These models came from SciKitLearn's machine learning library.

Results

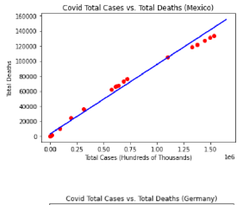
To answer the first research question, the multiple linear regression model found coefficients for each of the features listed: Total tests per thousand, positive test rate, stringency index, median age, GDP, extreme poverty rate, cardiovascular death rate, hospital beds per thousand, and life expectancy. Positive coefficients are positively correlated with total infections, and vice versa.

Feature	Coefficient
Total tests per thousand	9737
Positive rate	3496
Stringency index	-9058
Median age	7099
GDP per capita	-2080
Extreme poverty rate	430
Cardiovascular death rate	-3218
Hospital beds per thousand	-3359
Life expectancy	147

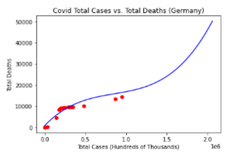
Simple regression was then used to predict total deaths based on total infections. This model was used globally as well as by country. The R^2 score for the global model was .946 which means the model has relatively low variance. However the mean squared error for this model was 204,424,213 due to many outliers. Simple regression improved greatly at the country level. Countries like with more linear infections vs deaths lines had much higher R^2 scores. Mexico, for example, had an R^2 score of .984 and a mean squared error of 36,246,091.



Polynomial regression was then used at the country level. This type of regression had improved R^2 score and mean squared errors for each country. Polynomial regression for countries with non-linear infections vs deaths lines also performed well. An example of this would be Germany (displayed on the right).



Finally when linear regression was used c tries on four features: GDP, stringency index, median age, and life expectancy. When the countries are organized by wealthiest continent, the GDP and stringency index of the wealthiest 3 continents are almost completely positively correlated with infection rate, while countries from the 3 poorest continents had GDP and stringency index that was negatively correlated with infections. When the countries were organized by infection rate, each of the features turned out to be negatively correlated with infection rate (with the Asian countries as outliers).



Feature Correlation with Infections of Countries Organized by Wealthiest Continents

Country	GDP	Stringency Index	Median Age	Life Expectancy
China	1.03	221	264	132
South Korea	-.14	217	138	69
Japan	0	2681	331	-331
US	0	11117	35334	0
Mexico	8.38	2878	-4294	0
Canada	0.54	6970	-555	227
France	3.46	8820	0	1772
Spain	10.47	10523	0	5363
Hungary	.63	2146	646	0
Iran	0	21160	0	2756
Iraq	3.62	-9764	0	463
Afghanistan	-2.71	-340	521	0
Brazil	-57.90	-42202	0	0
Argentina	-1.68	4574	1718	429
Uruguay	-.01	-21	3.8	-1.8
South Africa	5.29	-2378	0	0
Sudan	0	-233	-119	-30
Chad	-.14	-6	0	0

Features for Countries Organized by Infection Rate

Country	GDP	Stringency Index	Median Age	Life Expectancy
US	54225	71.76	38.3	78.6
France	38605	63.89	42.0	82.66
Hungary	26777	72.22	43.4	76.88
Spain	34272	78.7	45.5	83.56
Brazil	14103	60.65	33.5	75.88
Argentina	18933	79.17	31.9	76.67
Canada	44017	67.13	41.4	82.43
Uruguay	20551	70.37	35.6	77.91
South Africa	12294	47.22	47.22	64.13
Iran	19082	70.83	32.4	76.68
Iraq	15663	53.7	20.0	70.60
Mexico	17336	73.15	29.3	75.05
South Korea	35938	60.65	43.4	83.03
Japan	39002	53.24	48.2	84.63
Sudan	4466	28.70	19.7	65.31
Afghanistan	1803	12.04	18.6	64.83
Chad	1768	61.11	16.7	54.24
China	15308	78.24	38.7	76.91

Tools

Conclusions and Future Directions

In conclusion, each research question was answered effectively. Multiple linear regression was able to find coefficients for each of the nine features provided. Covid-19 deaths can be predicted accurately based on infection rate for simple linear regression and polynomial regression effectively. The models performed significantly better at a country level and with polynomial regression. Finally, linear regression was used with specific features for countries, and patterns were elicited effectively.

Findings for this project could be used in case of a future pandemic. Analysis of each of the features would also be helpful in curtailing or preventing a future pandemic. Finally, more features taken from different datasets could be analyzed.

Acknowledgments

Dr. Mohammed Aledhari, Assistant Professor of Computer Science

Contact Information

Noah Druss – noahdruss@gmail.com
Dr. Mohammed Aledhari - maledhar@kennesaw.edu

References

[1]"OWID/covid-19-data."[Online].Available:https://github.com/owid/covid-19-data

[2] K. Jolly,Machine Learning with scikit-learn Quick Start Guide: Classification, regression, and clustering techniques in Python.PacktPublishing Ltd, 2018.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion,O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg,J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot,and E. Duchesnay, "Scikit-learn: Machine learning in Python,"Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[4] A.Schneider,G.Hommel,andM.Blettner,"LinearRegression Analysis,"Deutsches Arzteblatt International, vol.107, no. 44, pp. 776–782, Nov. 2010. [Online]. Available:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2992018

[5] "Imf datamapper." [Online]. Available: https://www.imf.org/

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING